Visual Adversarial Examples Jailbreak Aligned Large Language Models

Year: 2023 | *Citations: 257* | *Authors: Xiangyu Qi, Kaixuan Huang, Ashwinee Panda*

Abstract

Warning: this paper contains data, prompts, and model outputs that are offensive in nature.
Recently, there has been a surge of interest in integrating vision into Large Language Models (LLMs), exemplified by Visual Language Models (VLMs) such as Flamingo and GPT-4. This paper sheds light on the security and safety implications of this trend. First, we underscore that the continuous and high-dimensional nature of the visual input makes it a weak link against adversarial attacks, representing an expanded attack surface of vision-integrated LLMs. Second, we highlight that the versatility of LLMs also presents visual attackers with a wider array of achievable adversarial objectives, extending the implications of security failures beyond mere misclassification. As an illustration, we present a case study in which we exploit visual adversarial examples to circumvent the safety guardrail of aligned LLMs with integrated vision. Intriguingly, we discover that a single visual adversarial example can universally jailbreak an aligned LLM, compelling it to heed a wide range of harmful instructions (that it otherwise would not) and generate harmful content that transcends the narrow scope of a `few-shot' derogatory corpus initially employed to optimize the adversarial example. Our study underscores the escalating adversarial risks associated with the pursuit of multimodality. Our findings also connect the long-studied adversarial vulnerabilities of neural networks to the nascent field of AI alignment. The presented attack suggests a fundamental adversarial challenge for AI alignment, especially in light of the emerging trend toward multimodality in frontier foundation models.